Streaming Long Video Understanding with Large Language Models

Year: 2024 | *Citations: 104* | *Authors: Rui Qian, Xiao-wen Dong, Pan Zhang*

Abstract

This paper presents VideoStreaming, an advanced vision-language large model (VLLM) for video understanding, that capably understands arbitrary-length video with a constant number of video tokens streamingly encoded and adaptively selected. The challenge of video understanding in the vision language area mainly lies in the significant computational burden caused by the great number of tokens extracted from long videos. Previous works rely on sparse sampling or frame compression to reduce tokens. However, such approaches either disregard temporal information in a long time span or sacrifice spatial details, resulting in flawed compression. To address these limitations, our VideoStreaming has two core designs: Memory-Propagated Streaming Encoding and Adaptive Memory Selection. The Memory-Propagated Streaming Encoding architecture segments long videos into short clips and sequentially encodes each clip with a propagated memory. In each iteration, we utilize the encoded results of the preceding clip as historical memory, which is integrated with the current clip to distill a condensed representation that encapsulates the video content up to the current timestamp. After the encoding process, the Adaptive Memory Selection strategy selects a constant number of question-related memories from all the historical memories and feeds them into the LLM to generate informative responses. The question-related selection reduces redundancy within the memories, enabling efficient and precise video understanding. Meanwhile, the disentangled video extraction and reasoning design allows the LLM to answer different questions about a video by directly selecting corresponding memories, without the need to encode the whole video for each question. Our model achieves superior performance and higher efficiency on long video benchmarks, showcasing precise temporal comprehension for detailed question answering.